Lexical Explorer: extending access to the Database for Spoken German for user-specific purposes

Lemmenmeier-Batinić, Dolores

Abstract: This paper presents Lexical Explorer,2 a tool that allows interactive browsing and filtering of quantitative corpus information. It further describes how this tool can be used to support linguistic work on corpora of spoken German. By using Lexical Explorer, users can analyse quantitative corpus data by interacting with frequency tables and obtaining customised word profiles of word distribution across word form variation, co-occurrences and metadata. Interaction with corpus examples of particular corpus counts is also enabled. Lexical Explorer was developed as a prototype for user-specific corpus access and is aimed at researchers of German lexicon in spoken interaction. Although Lexical Explorer was developed on the basis of two small speech corpora of the German language, the underlying principle of this tool can be easily adapted to other corpora and other user groups. Moreover, the tool can be used to gain insights into the corpus structure as well as to study and verify corpus content in a transparent and user-friendly way.

DOI: https://doi.org/10.3366/cor.2020.0185

Dolores Lemmenmeier-Batinić
dolores.lemmenmeier@uzh.ch

Lexical Explorer: Extending access to the Database for Spoken German for user-specific purposes

Abstract

This paper presents the Lexical Explorer[1], a tool that allows interactive browsing and filtering of quantitative corpus information. It further describes how this tool can be used to support linguistic work on corpora of spoken German. By using the Lexical Explorer, users can analyse quantitative corpus data by interacting with frequency tables and obtaining customised word profiles of word distribution across word form variation, co-occurrences and metadata. Interaction with corpus examples of particular corpus counts is also enabled. The Lexical Explorer was developed as a prototype for user-specific corpus access and is aimed at researchers of German lexicon in spoken interaction. Although the Lexical Explorer was developed on the basis of two small speech corpora of the German language, the underlying principle of this tool can be easily adapted to other corpora and other user groups. Moreover, the tool can be used to gain insights into the corpus structure as well as to study and verify corpus content in a transparent and user-friendly way.

Keywords: spoken German, talk-in-interaction, user group differentiation

1. Introduction

This paper presents an interface for browsing and filtering the quantitative data of two corpora situated in the Database for Spoken German (DGD): FOLK (Forschungs- und Lehrkorpus Gesprochenes Deutsch [German Research and Teaching Corpus of Spoken German], Schmidt 2014a) and GeWiss (Gesprochene Wissenschaftssprache [Spoken academic language], Fandrych, Meißner and Wallner 2017). I extended access to the corpora by building an interface that provides insights into the corpus content from a quantitative perspective by allowing the user to compare frequency distributions of particular phenomena without having to first query the corpus.

Developing the Lexical Explorer constituted a first step towards designing an architecture for spoken corpus platforms that allows for a more flexible differentiation of corpus access according to the needs of different user groups. The differentiation of corpus access for different user groups was proposed by Anthony (2013) as a feature of next-generation corpus tools modelled according to the Model View Controller (MVC) architecture. Extending corpus access according to these principles is currently being elaborated in detail in the project "Zugang zu multimodalen Korpora gesprochener Sprache: Vernetzung und zielgruppenspezifische Ausdifferenzierung" (ZuMult) at the Institute for the German Language in Mannheim.[2]

From the perspective of user group differentiation, the Lexical Explorer is a prototype for corpus access dedicated primarily to researchers of spoken German lexicon. It was originally developed in order to enhance the creation of a corpus-based lexicographic resource for spoken German, aimed at researchers of spoken standard German in talk-in-interaction

[1] http://www.owid.de/lexex (14.07.2018).
[2] http://zumult.org/ (21.09.2018).

(LeGeDe – Lexik des gesprochenen Deutsch [Lexicon of spoken German]; Möhrs et al. 2017). The intention is to include the Lexical Explorer in the infrastructure of the lexical resource developed in the LeGeDe project.

To assist the creation of the LeGeDe lexical resource, I performed a series of calculations, which were then used to gain insights into quantitative word distributions in FOLK. In order to make this information available to a broader academic audience, I implemented the option to search and filter the results of these calculations in a web interface. Both analysis and corpus access are based on lemmatised lexical units.[3] The tool delivers the data that are more difficult or not possible to obtain by querying and quantifying the corpus in the DGD. Whenever possible, the lexical units presented in the Lexical Explorer are connected to the concordance view in the DGD.

To perform a quantitative analysis of a corpus of spoken language consisting of transcriptions of talk-in-interaction, it was necessary to first consider the respective aspects that are not encountered when working with corpora of written language and subsequently design appropriate pre-processing tasks. For instance, I had to determine how to deal with non-verbal elements, pauses and uncertainties in transcriptions in the pre-processing stage. Moreover, working with a corpus of spoken language made it possible to detect and quantify phenomena that are seldom encountered in written corpora but are recurrent in the talk-in-interaction, such as repetitions and disfluencies, and various (transcribed) pronunciation variants. In addition, the Lexical Explorer also enables users to investigate the distribution of lexical units across meta-information, e.g., gender, communication events and interaction categories (see Section 3.2.5). These phenomena cannot be extracted by using the corpus tools designed primarily for written language corpora, such as SketchEngine (Kilgarriff 2014), AntConc (Anthony 2018), WordSmith Tools (Scott 2016), etc. However, although the quantitative analysis of the corpus had to be elaborated in order to meet the challenges of a corpus of spoken language, the online presentation's underlying principle as well as the functionalities of the Lexical Explorer are not specific for spoken language data, and can be easily transferred to other corpora and adapted to the needs of other user groups.

In the next sections, I will present the DGD database and the corpora that were considered for the corpus view presented in the Lexical Explorer (Section 2), the considerations regarding the creation of the Lexical Explorer, data pre-processing, calculation and presentation (Section 3), and practical examples of use of the Lexical Explorer (Section 4). For reasons of consistency, all the examples in the paper are gathered from the FOLK corpus found in the DGD version 2.8.

2. FOLK, GeWiss and the Database for Spoken German

FOLK (Schmidt 2014a) is a corpus of naturally occurring verbal interaction that is sufficiently large and diverse to support various quantitative and qualitative research approaches. FOLK contains richly annotated transcriptions (orthographic normalisation, lemmatisation, part-of-speech tagging, non-verbal elements) and metadata (speaker and event information). The

[3] In this paper, the term "lexical unit" is used to denote the transcribed words and multiword expressions as well as their normalised and lemmatised counterparts.

corpus is divided into contributions with one contribution being a sequence of words that is not interrupted by a pause longer than 0.2 seconds. The transcription is performed according to the GAT conventions (Selting et al. 2009).

GeWiss (Fandrych, Meißner and Wallner 2017) is a comparative corpus for spoken academic language containing transcripts in German, Polish, English, and Italian (for this paper, I only worked with the German version). The academic language in GeWiss is represented by student and expert presentations and conversations during exams. As in the case of FOLK, the audio-data in GeWiss is transcribed with GAT conventions.

FOLK and (German) GeWiss are accessible through the DGD (Schmidt 2014b). The DGD enables online browsing of transcripts and audio/video files, as well as Key-Word-In-Context (KWIC) searches. It also shows the distributions of the searched keyword(s). In addition to the positional and metadata search, there is also the possibility to examine the context by querying each neighbour token individually. The DGD content can be accessed through full text search as well as through concordances (Figure 1). After the keyword search, the user has the option to quantify the corpus search on different levels, as shown in Figure 2. The quantification is also provided for metadata (not shown in Figure 2 for space reasons). There are, however, no direct links from quantified data to corpus examples. A new search is required for further examination.

Figure 1: KWIC concordance with transcript excerpt for the lemma *gut* (good) in the DGD

Figure 2: An excerpt of the KWIC quantification for the lemma *gut* (good) showing the corpus overview and the counts regarding word annotation layers in FOLK (transcription, orthographic normalisation, lemmas and part-of-speech tags)

Although the functionalities implemented in the DGD already proved to be adequate to conduct diverse studies in the field of interactional linguistics (Deppermann and Helmer 2013; Deppermann and Schmidt 2014; Zeschel and Proske 2015; Kaiser 2017), the need to extend the corpus with quantitative information and diversified access became apparent in the process of creating the lexical resource of spoken German (Möhrs et al. 2017). In order to facilitate the selection and analysis of salient terms of spoken German, we[4] needed to be able to examine the quantitative data in a corpus-driven manner and to create an intermediate step that could direct the corpus search to more specific queries. The solution was to create a platform on which the quantitative corpus data could be browsed, sorted, filtered and linked to the corresponding concordances in the corpus.

3. Extending access to DGD through the Lexical Explorer

My aim for the extension of the quantitative DGD data was to enable corpus-driven as well as inductive searches of corpus counts. Moreover, I wanted the option to connect different types of corpus counts for the same target word, so that the user would be able to query and examine different quantitative information simultaneously. In order to put this into practice, I calculated corpus distributions of different investigated phenomena (word forms, co-occurrences, metadata, etc.), stored them in form of tables and presented them online in a web

[4] With „we“ I refer to the team working on the project „The Lexicon of Spoken German“ (LeGeDe: Lexicon des Gesprochenen Deutsch, Möhrs et al. 2017).

application (see Section 3.3). I enabled the option to browse and query all tables individually and collectively when searching for a keyword lemma. In order to connect the quantitative data presented in the Lexical Explorer to the corpus examples in the DGD, the external query parameters redirecting the user to the corpus search are sent from the Lexical Explorer in the DGD. Hence, external access to the DGD is performed by first preselecting and processing the corpus data addressing one specific user group, and second, by connecting that corpus data to the corpus examples stored in the corpus database (DGD). Needless to say, prior to the second step, the corpus query infrastructure has to be designed in such a way as to accept query parameters from an external source.[5]

The absence of adequate segmentation[6] (the transcripts not being segmented by linguistic criteria but by pause longer than 0.2 s) and the high frequency of disfluencies in spoken language are only a few examples of the challenges that had to be considered in the quantification process.

Currently, the set of tables presented in the Lexical Explorer is divided into five sections: word units, co-occurrences, repetitions, metadata, and event keywords. The section *Word units* covers the distribution of word transcriptions, the comparison of lemma frequencies in spoken and written German, and the distribution of separable particle verbs. The section *Co-occurrences* contains bi- and trigrams as well as co-occurrences in the context of ±5 tokens. The section *Repetitions* presents the quantification of lexical repetitions that comprise disfluencies in spoken language interaction as well as repeated non-grammatical elements, such as *ja ja*. Information about gender and interactional categories is stored in the section *Metadata.* The section *Event keywords* contains most commonly used lemmas in a particular event in comparison to all other events in the corpus.

3.1. Data preparation

Depending on which information was relevant for the individual table, the lemma layer and, where needed, all annotation layers (transcription, orthographic normalisation, lemma, parts-of-speech) were considered. For instance, in the calculation of the bigrams, I did not consider the transcription layer because it was beyond our scope to investigate pronunciation variations in the case of co-occurrences. However, this layer was included in the tokens table, which offers an overview of all transcription variants.

The categories of the part-of-speech tagset STTS 2.0, which was originally developed for FOLK (Westpfahl 2014), were mapped to their simplified part-of-speech tags that came to constitute an additional annotation layer considered in the calculations. These categories are AB (abruptions), ADJ (adjectives), ADV (adverbs), AP (prepositions and adpositions), ART (articles), CARD (cardinals), FM (foreign-language units), KO (conjunctions), N (nouns), NG (non-grammatical elements/sentence independent elements), ORD (ordinals), P (pronouns), PTK (particles), SE (sentence-external elements), UI (unintelligible instances), and V (verbs).

[5] At the time of writing, the external query search for searching the co-occurrences in the DGD was still being developed.
[6] The segmentation of FOLK is currently being elaborated in the SegCor project at the Institute for the German Language in Mannheim: http://www1.ids-mannheim.de/prag/muendlichekorpora/segcor.html, 11.05.2018.

These categories can be considered an adaptation of the universal parts-of-speech proposed by Petrov et al. (2011) to spoken German.

The transcriptions in FOLK contain several symbols for non-standard variants and peculiarities of spoken language (Table 1). These phenomena required special treatment in terms of the calculations of co-occurrences (see Section 3.2.3).

Table 1: Symbols for disfluencies in FOLK

Uncertain transcriptions, as well as alternative suggestions for uncertain transcriptions are marked in FOLK. However, in the calculations for the Lexical Explorer only uncertain transcriptions but not the alternatives were considered. Contribution-internal pauses and non-verbal elements were ignored in each calculation.

3.2. Data calculation

3.2.1. Distribution of word forms

To enable an easier exploration of transcription (i.e. pronunciation) variants of particular word forms in the corpora, the table *Tokens* provides a structured representation of each transcription along with all its annotation layers, including lemma, orthographic normalisation, transcription, part-of-speech, and simplified part-of-speech. For the creation of this table, the disfluencies shown in Table 1 were included. In addition to studying word distribution on different annotation levels, this representation can be used from the developer side as support for finding transcription and annotation inconsistencies in the corpus as well as unintelligible words, abruptions, and other types of disfluencies.

3.2.2. Frequency comparison between spoken corpora and DEREKO

The table *Study corpus vs. DeReKo* provides an overview of lemma frequencies in the spoken corpora in comparison to the German Reference Corpus of written language (Deutsches Referenzkorpus, DEREKO, cf. Kupietz and Keibel 2009), amounting to approximately 30 billion tokens (Version 2017-I). The comparison is focused on the corpora of spoken language (for instance, the table shows the frequencies of all lemmas that occur in FOLK to their frequencies in DEREKO, not the other way around), and is based on different measures for corpus comparison that have been taken into account for the purpose of detecting lexical peculiarities of spoken language (Möhrs et al. 2017). By default, the search result for the corpus comparison shows the values of frequency classes of each lemma in both corpora ("Häufigkeitsklasse" HK; Keibel 2008, 2009) as well as the differences in frequency classes. The most common word in a corpus has the frequency class 0, the word(s) being approximately half as frequent as the most frequent word have the frequency class 1, the words being approximately half as frequent as those in the class 1 have the frequency class 2, etc. The comparison of lemma frequency classes in FOLK and DEREKO presented in this overview was used for the process of selecting the headword candidates for the lexical resource of spoken German (Meliss et al. 2018).

In order to guarantee transparency and reliability of the results, along with the calculations for statistical measures (frequency classes, log ratio, odds ratio, risk ratio, log likelihood, chi

square), the number of absolute occurrences for each corpus (marked as *freq* in the table column headers) is provided.

Since in the lemma comparison we were not interested in proper names, numbers and words that are only used in one transcript or by only one speaker, I marked them as outliers and set the option to filter them in the table setting (column *Filter*). Furthermore, I added (simplified) parts-of-speech for each lemma in order to facilitate filtering according to word classes.

3.2.3. Co-occurrences

I implemented an overview of bigrams, trigrams and co-occurrences within the context of ±5 tokens. The calculation of co-occurrences within the context of ±5 tokens was only performed for the lemma-layer. For each lemma (*Lemma 1*), I calculated how often it occurs in the same context window with any other lemma in the corpus (*Lemma 2*). In the case of bigrams and trigrams, the layers lemma, orthographic normalisation and simplified parts-of-speech are considered in order to discriminate between different homographs with different lemmas and parts of speech. This enables a fine-grained exploration of bigrams, such as the possibility to differentiate between left or right contexts of particular lexical units.

The calculation of co-occurrences is performed for all the tokens within one contribution. Although segmentation of contributions in FOLK and GeWiss is not yet based on syntactic and prosodic criteria, the contributions are a more accurate representation of sentence equivalences in spoken language than, for instance, speaker turns. Calculating the co-occurrences within speaker turns might result in syntactically independent tokens represented in the same bigrams and trigrams.

Beginnings and ends of contributions are included in n-gram counts and marked as <.>. Although they are not technically part of bigrams and trigrams in the traditional sense, I considered them in the representation because they provide valuable contextual information, especially for research of talk-in-interaction, in which attention is often payed to positional information (contribution beginnings and endings, stand-alone units, etc.). For instance, by setting the <.> in the first and last position of a trigram, it is possible to inspect the most frequent tokens that occur as stand-alone items in a contribution. Short pauses and non-verbal elements within a contribution (cough, breathing, etc.) were filtered out for the calculation of co-occurrences, as well as unintelligible words, stutters, and placeholders for abbreviations and abruptions. For instance, in the example of an abruption “phonolo <pause> phonologischen” (Table 1), the normalisation “phonologischen %” did not count as a co-occurrence of the following token. The counts of co-occurrences are given in terms of absolute frequencies and log likelihood ratio.

Although GeWiss and FOLK are not yet parsed on the syntactic level, it was possible to extract one particular kind of syntactically dependent co-occurrence: the separable particle verbs. Relying on the reconstruction principle of particle verbs within contributions (Batinić and Schmidt 2018), I calculated the distribution of the particle verbs when they occur together and separately, and integrated the list of all separable verbs with their respective counts. For each example, a direct link to the DGD concordances is implemented. This view is the only

possibility to access *all* separable verbs in the two corpora, and not only those that are written together.

3.2.4. Repetitions

I defined a repetition of a lexical unit to be the exact repetition of a transcribed word that also has an identical orthographic normalisation, part-of-speech tag and lemma to the previous unit. Accordingly, phenomena such as self-corrections realised as slightly modified repeats (Table 2), or reductions in pronunciation (Table 3) were not detected as repetitions. I was primarily interested in imminent and exact word repetitions that could assist the study of the effect of doubling responses in talk-in-interaction, such as *ja ja* or *nein nein*. However, since the algorithm was applied to all tokens in the corpus and not only to the selected response tokens, all imminent lexical repeats were detected. The principle of detection was the following: for each lexical unit in a contribution, check if the next one and all its annotations are identical to the previous one. If the condition is met, repeat until this condition is no longer met.

As in the case of co-occurrences, lexical repetitions were detected within one-contribution boundaries, since every new contribution is potentially syntactically and thematically different from the previous one. I did not consider a sequence to be repeated if any type of hiatus other than non-phonological instances such as short pauses and breathing was in-between.

Table 2: Self correction/abruption of *könnte* (could), FOLK_E_00069

Table 3: Reduction in pronunciation of *mach mal* to *ma ma* in the contribution *ma ma scheiß ding au,* FOLK_E_00030

Once all the repetitions were detected, I calculated the number of all repetitions for each lexical unit overall and the absolute frequency of the respective lexical units in the corpus. In addition, I discriminated between the repetitions of two, three, four, and five or more than five lexical units (Table 4). For the calculation and visual presentation, I relied on three annotation layers: lemma, orthographic normalisation and simplified part-of-speech, which were also basic units for the calculation of co-occurrences. Although repetition of the word transcription was fundamental for the detection of repetitions, transcription variations were not included in the table presentation of the data, since they would affect the clarity of the presentation. However, for each repetition, direct links to corpus examples containing all variations of word transcriptions are implemented. The structure of the resulting table is based on the quantification of reduplications presented in Deppermann and Helmer (2013), which previously had to be created manually by querying each individual table item on the DGD surface.

Table 4: Excerpt of the table containing counts of exact one-word repetitions in FOLK

The same procedure was repeated for two-word repetitions. As with the one-word repetitions, the two-word repetitions had to have identical transcriptions, orthographic normalisations, lemmas and parts-of-speech in order to be detected as such. Similarly, they also had to be in

direct continuation, i.e. if there was any type of hiatus in between (except for short pauses and non-verbal elements) they were not considered imminent lexical repetitions.

3.2.5. Metadata

I integrated various statistical measures that represent the extent to which lemma frequencies differ in different gender groups. As in the comparison of spoken corpora with DEREKO (see Section 3.2.2), parts-of-speech for each lemma as well as the possibility to filter out the outliers were added to the table.

In addition to gender information, I calculated the distribution of lemmas across the interactional categories (private, non-private/non-public, public and other) that were introduced in order to study the lexicon of spoken language across different interactional categories (Möhrs et al. 2017). The category *private* comprises events of everyday conversations among friends and family, the category *non-private/non-public* contains events of interaction at school, university or at the workplace, and the category *public* refers to mediation talks and panel discussions. For each category, the absolute (*Abs*) and relative counts (*Rel*) of lemmas occurring in each category are presented.

3.2.6. Event keywords

For each communication event presented in the corpus, I extracted twenty lemmas with the highest term frequency-inverse document frequency index (tf-idf). The extraction of most highly ranked tf-idf lemmas enabled an overview of the keywords for each event as well as an overview of all the events in which one particular lemma is encountered more often than usual. The event ids and descriptions were given for each event. An example of tf-idf keywords for the event bible study group (FOLK_E_00193) is presented in Table 5. Among ten lemmas with the highest tf-idf score presented in Table 5, it is possible to observe typical keywords commonly used in religious contexts, such as *Sünde* (sin), *Geist* (spirit), *Gott* (God), etc.

Table 5: Tf-idf keywords of the event bible study group (FOLK_E_00193)

3.3. Data presentation and functionalities

To enable interactive querying of the tables on a web interface, I implemented server-side processing of the jQuery plug-in DataTables[7] with the Oracle database by using the Python Framework CherryPy[8]. The table content is fetched from the database by using AJAX requests. I enriched all the tables with features such as searches for each column and added wildcard searches in order to allow for more complex queries (Table 6).

Table 6: Wildcards for querying the Lexical Explorer

In contrast to the word quantifications shown in Figure 2 (see Section 2), the differences in the distribution of different token and parts-of-speech occurrences are not only structured,

[7] https://datatables.net/ (28.02.2018).
[8] http://cherrypy.org/ (28.02.2018).

sorted and visible at one glance in the Lexical Explorer, but there is also a direct corpus link for each example (Figure 3, marked in bold). For instance, when exploring the occurrences of the form *gut* as a lemma, the user can observe the distribution of the parts-of-speech for each form and use a link with the predefined specific query of their interest.

Figure 3: An excerpt of the lemma search of *gut* (good) showing the token frequency at each annotation level

To improve the clarity of the presentation, we set default column visibility (Figure 4) and allow the users to adapt it to their own preferences. Each table and column can be sorted, filtered and browsed individually, but since each table is provided with a lemma, there is also a possibility to filter all the tables at once by using the lemma search function (*Filter all tables by lemma (keyword)*). The result of the lemma search is a set of filtered tables that act as an automatically generated word profile. For most tables, a direct link to the DGD corpus examples is implemented and can be queried without leaving the Lexical Explorer interface, provided that the user has a DGD account.

Figure 4: Default setting of columns for the bigrams table: the orthographic normalisations of the first (*Norm 1*) and second (*Norm 2*) bigram member and their absolute frequency (*Freq*)

Each table displays 10 to 100 results per page (user-defined). The results can be exported to CSV format. If the link to the corpus examples is available, the corresponding entry is marked in bold and linked to the DGD. The lemma search for all tables per default filters the first lemma column (for example, in the case of bigrams, it automatically filters the first element), but can also be adapted for each position (click on *Filter keyword in {x} position*).

4. Using the query interface

In the following section, I will show how the Lexical Explorer can be used to gain insights into the frequency distribution of a particular lemma. We do not make prior assumptions about the investigated phenomenon and let the data guide us in the querying and filtering process.

4.1. *Gott*

Expressions of religious origin are widely used in everyday conversation in spoken German. However, although the interest of researchers in describing different kinds of pragmatic expressions has increased in the last years (Aijmer 1996; Schourup 1999; Norrick 2009), the variation, distribution and pragmatic functions of "religious" expressions such as *Gott* (God), *Jesus* (Jesus), and *Maria* (Mary), etc. have not been studied to a great extent in spoken German. In the following paragraphs, I explain how the Lexical Explorer can assist the process of studying the distribution of the lemma *Gott* in FOLK.

In order to study the occurrences of *Gott*, let us first observe the quantitative distribution of the lemma *Gott* by using the Lexical Explorer and considering the top 10-20 hits of each table. The table in Figure 5 shows that *Gott* occurs either tagged as N (noun) or as a non-grammatical-element (NG), more precisely, NGIRR, which stands for interjections, response signals and backchannel behaviour. After inspecting all the records (total: 14), we observe

that *Gott* is annotated as NGIRR more than twice as often as in the noun sense (557:270). For further study of particular examples, a user can click on a word form (in bold) and will be redirected to the respective corpus examples.

Figure 5: Occurrences of the word forms with the lemma *Gott* at each annotation level (*Freq*: frequency in FOLK)

The table *Study corpus vs. DeReKo* (in Figure 6: *FOLK vs. DeReKo*) reveals that *Gott* is used more frequently in FOLK than in DeReKo: the frequency class of *Gott* in FOLK is 7, meaning that *Gott* is located in the seventh most frequent word group in FOLK (cf. Keibel 2008, 2009). In DeReKo the frequency class of *Gott* amounts to 10. The prevalence of the lemma *Gott* in FOLK can be attributed to the fact that interjections are much more common in spoken than in written mode. However, it may also suggest that FOLK contains events in which the literal meaning of the word *Gott* is used more frequently than one may expect.

Figure 6: Excerpt from the table *Study corpus vs. DeReKo* (*HK Diff* stands for the difference of frequency classes

Since *Gott* as a non-grammatical element (interjection) occurs more frequently then the usage of *Gott* as a noun in the FOLK corpus, most frequent collocates also relate to the NG meaning (*oh Gott, ach Gott, mein Gott*; Figure 7). In addition, the output of the table *Two-word repetitions* suggests that there is a tendency to repeat the most common collocates of *Gott*: *oh Gott* and *ach Gott* (Figure 8).

Figure 7: Excerpt from the table *Bigrams* showing the most common bigrams of *Gott* in FOLK in left context

Figure 8: Distribution of repetitions of *oh Gott* and *ach Gott* in FOLK: *SUM_REPS* refers to the sum of the repeated tokens; *R{x}* refers to the number of times the expression was repeated[9]

When searching for other interjections that appear before *Gott,* we query the tag NG in the left context of the lemma *Gott* (enable the full column visibility and then click on *PoS 1* to show the column containing the parts-of-speech of the lemma in left context – *Lemma 1*). The filtered table reveals that *Gott* also occurs, although much less frequently, in the context of other religious expressions such as *Jesus* or *Jesses,* the latter being a spoken language variant for *Jesus* in exclamations such as *Jesses Gott.* The <.> before *Gott* in Figure 7 shows that *Gott* also frequently occurs without exclamation particles or any other lexical units in left context (84 occurrences). In order to verify how often it occurs as a stand-alone item, we search for *Gott* in one-word contributions: <.> Gott <.> (Figure 9). The result shows that *Gott* is used 27 times as a stand-alone expression.

[9] There are 6 examples in the corpus in which *oh Gott* is repeated twice, meaning that we find 12 instances of *oh Gott* if we look at these examples (6x2). Two times *oh Gott* is repeated as a sequence of three *oh Gott oh Gott oh Gott* (2x3). Hence, we have a total of 18 *oh Gott* that form part of a repetition (*SUM_REPS*). This number can be used to calculate the proportion of the repetitions in relation to the overall occurrence of the searched word or multiword unit.

Figure 9: Excerpt from the table *Trigrams* showing *Gott* preceded and succeeded by a long stop <.>, standing for the contribution boundary

After having considered the peculiarities of *Gott* in the interjection sense, let us investigate the collocates of the noun meaning of *Gott.* We search for all the adjectives describing *Gott* by querying the tag ADJ in the part-of-speech column of the left context (Figure 10). The filtered table suggests that the adjectives that co-occur most commonly together with the noun meaning of *Gott* in FOLK are, among others, *lieber* (dear), *mickriger* (puny)*, böser* (evil), *gewordener* (who became)*,* and *lebendiger* (living).

Figure 10: The most common bigrams with the word form *Gott* as second element if it is tagged as a noun (N) and if it is preceded by an adjective (ADJ)

The proportion of males and females is almost evenly distributed in FOLK (DGD version 2.8: 948586 tokens spoken by males, 980190 by females). According to the information in the *Gender* table, female speakers use the lemma *Gott* almost three times as often as male speakers (600:213). Since this table does not discriminate between noun and interjection meaning, we can check this information again by searching for the occurrences of the lemma *Gott* once as an interjection and once as a noun in the corpus examples and by quantifying the occurrences according to gender. After this step, we observe that the lemma *Gott* tagged either as an interjection or as a noun is used more often by female speakers (415:127; 183:85).

Since the corpus is relatively small in size, it is very important to pay attention to the event and speaker metadata when interpreting corpus counts, since the searched phenomena might occur very frequently in one transcript or by the same speaker. The bigram *mickriger Gott* (Figure 10) for instance is used only in one event (FOLK_E_00204) when referring to a quote *puny God* from the film Avengers. In addition, the examination of the table *Tf-idf keywords* reveals that the events in which the lemma *Gott* is the most frequent are the bible study group (FOLK_E_00193) and the student everyday-conversations (FOLK_E_00048), and that in both transcripts, there are only three female speakers each. The question whether the difference in the frequency of *Gott* between the male and female speakers is due to gender differences or whether it can be explained by the prevalence of female speakers in particular communicative events is left to further corpus examination by researchers interested in the gender question.

Lastly, the quantitative analysis of *Gott* can also be approached by exploring the interaction categories in which *Gott* appears, which are represented in the table *Categories* (Figure 11), and which show that the lemma *Gott* in FOLK occurs most often in private conversations (536 occurrences) and is least common in public conversations (11 occurrences).

Figure 11: Distribution of the lemma *Gott* in the categories private, non-private/non-public and public.

Conclusion

I have described the Lexical Explorer, a tool is used for the exploration of quantitative information in a corpus of spoken language. The Lexical Explorer is available via web-interface which is connected to the corpus counts as well as to the corpus examples with the corpus login only being required when accessing the corpus examples. Although it was developed for researchers focussing on the lexicon of spoken German in interaction, the data presented in the Lexical Explorer could also be relevant for other linguistic disciplines, such as interactional linguistics, corpus linguistics, lexicography and sociolinguistics. Further, the principle of filtering different tables simultaneously in order to simulate word profile views can be useful in other disciplines as well.

Working with richly annotated corpora such as FOLK and GeWiss allows for a high degree of flexibility with regard to the level of detail in querying the quantitative corpus data. Thanks to the annotation levels provided for each word, I could choose how granular I wanted the calculations to be and whether they should be based on lemma, orthographic normalisation, transcribed words or other features. In order to prevent annotation overload and to assure clarity of the data, I defined the level of granularity for each table separately, which required careful pre-processing of the corpus data and prior planning of the visual presentation.

Given that the data in the Lexical Explorer is transparent and in most cases verifiable through direct corpus links, I believe that this tool can support linguistic research at various levels. It can be used to study word variation at the transcription level, frequent word patterns, distribution of co-occurrences, repetitions, and particle verbs. It can additionally be used to investigate the distribution of particular morphosyntactic categories in spoken language and to explore the output of statistical measures commonly used in corpus linguistics. Lastly, this tool is also useful for corpus-creators, since it enables simple tracking of tagging, normalisation and lemmatisation (in)consistencies.

Bibliography

Aijmer, K. 1996. *English Discourse Particles: Evidence from a Corpus.* Amsterdam/Philadelphia: John Benjamins.

Anthony, L. 2013. 'A critical look at software tools in corpus linguistics' in *Linguistic Research* 30(2), pp. 141-161.

Anthony, L. 2018. AntConc (Version 3.5.7) [Computer Software]. Available from http://www.antlab.sci.waseda.ac.jp/ (Last accessed at 31.05.2018).

Batinić, D. and Schmidt, T. 2018. 'Reconstruction of Separable Particle Verbs in a Corpus of Spoken German' in Rehm G., Declerck T. (eds.): *Language Technologies for the Challenges of the Digital Age.* GSCL 2017. Lecture Notes in Computer Science, volume 10713. Springer, Cham, pp. 3-9.

Deppermann, A. and Schmidt, T. 2014. 'Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik - Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2)' in Domke, C. , Gansel, C. (eds.): *Korpora in der Linguistik - Perspektiven und Positionen zu Daten und Datenerhebung* [=Mitteilungen des Deutschen Germanistenverbandes 1/2014], pp. 4-17.

Deppermann, A. and Helmer, H. 2013. 'Standard des gesprochenen Deutsch: Begriff, methodische Zugänge und Phänomene aus interaktionslinguistischer Sicht' in Hagemann, J., Klein, W. P.,

Staffeldt, S. (eds.): *Pragmatischer Standard*. Tübingen: Stauffenburg. (Stauffenburg Linguistik 73), pp. 111-141.

Fandrych, C., Meißner, C. and Wallner, F. (eds.) 2017. *Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora.* Tübingen: Stauffenburg.

Kaiser, J. 2017. 'Reformulierungsindikatoren im gesprochenen Deutsch: Die Benutzung der Ressourcen DGD und FOLK für gesprächsanalytische Zwecke', *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 17 (2016), pp. 196-230.

Keibel, H. 2008, 2009. *Mathematische Häufigkeitsmaße in der Korpuslinguistik: Eigenschaften und Verwendung.* Mannheim: Institut für Deutsche Sprache (http://www.ids-mannheim.de/kl/dokumente/freqMeasures.html. Last accessed at 30.05.2018).

Kilgarriff, A., Baisa V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. 2014. 'The Sketch Engine: ten years on', *Lexicography* 1 pp.7-36.

Kupietz, M. and Keibel, H. 2009. 'The Mannheim German Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research' in *Working Papers in Corpus-based Linguistics and Language Education, No. 3*. Tokyo: Tokyo University of Foreign Studies (TUFS), 53-59.

Meliss, M., Möhrs, C., Batinić, D. and Perkuhn, R. 2018. 'Creating a list of headwords for a lexical resource of spoken German', to appear in *Proceedings of the Euralex Conference*. Ljubljana, July 17-21, 2018.

Möhrs, C., Meliss, M. and Batinić, D. 2017. 'LeGeDe – Towards a corpus-based lexical resource of spoken German' in Kosem, I., Tiberius, C., Jakubíček, M., Kallas, J., Krek, S., Baisa, V. (eds.): *Proceedings: Electronic Lexicography in the 21st Century.* Leiden, pp. 281-298.

Norrick, N. R. 2009. 'Pragmatic markers: Introduction', *Journal of Pragmatics* 41/5, pp. 863-865.

Petrov, S., Das, D. and McDonald, R. 2011. 'A universal part-of-speech tagset'. Available at http://arxiv.org/pdf/1104.2086.pdf (Last accessed at 07.05.2018).

Schmidt, T. 2014a. The Research and Teaching Corpus of Spoken German - FOLK in *Proceedings of LREC'14*, Reykjavik, Island: ELRA, pp. 383-387.

Schmidt, T. 2014b. 'The Database for Spoken German - DGD2' in *Proceedings of LREC'14*, Reykjavik, Island: ELRA, pp. 1451-1457.

Schourup, Lawrence. 1999. 'Discourse markers', *Lingua* 107/3–4, pp. 227-265.

Scott, Mike. 2016. *WordSmith Tools,* version 7, Stroud: Lexical Analysis Software.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S. and Hartung, M. 2009. 'Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)', *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion* 10, pp. 353-402.

Westpfahl, S. 2014. 'STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data' in Levin L. and Stede, M. (eds.): *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop.* Dublin, Irland: Association for Computational Linguistics and Dublin City University, pp. 1-10.

Zeschel, A. and Proske, N. 2015. 'Usage-based linguistics and conversational interaction. A case study of German motion verbs' in *Yearbook of the German Cognitive Linguistics Association* 3, pp. 123-144.

Figures

Figure 1: KWIC Concordance with transcript excerpt for the lemma *gut* (good) in the DGD

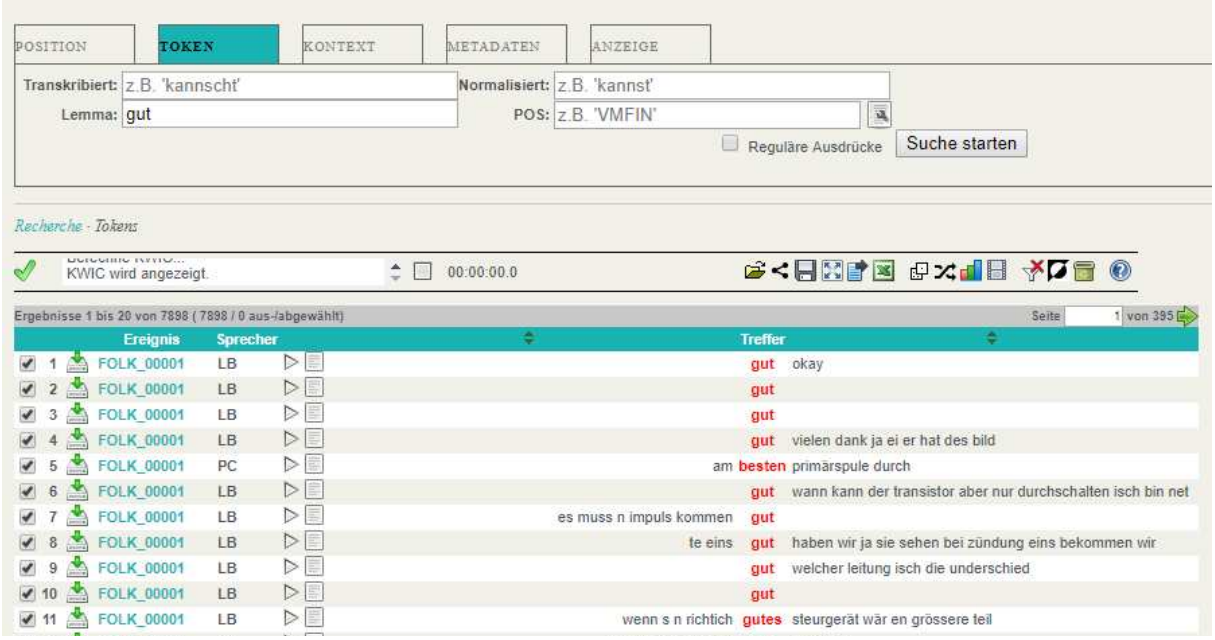
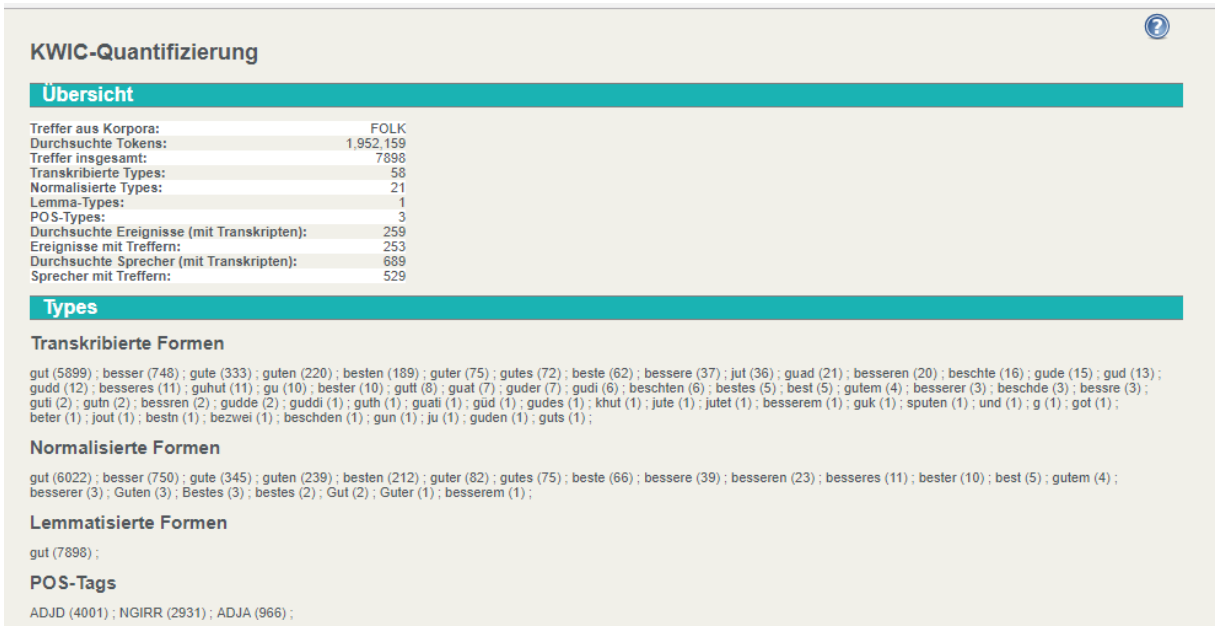


Figure 2: An excerpt of the KWIC quantification for the lemma *gut* (good) showing the corpus overview and the counts regarding word annotation layers in FOLK (transcription, orthographic normalisation, lemmas and the part-of-speech tags)



KWIC-Quantifizierung

Übersicht

Treffer aus Korpora:	FOLK
Durchsuchte Tokens:	1,952,159
Treffer insgesamt:	7898
Transkribierte Types:	58
Normalisierte Types:	21
Lemma-Types:	1
POS-Types:	3
Durchsuchte Ereignisse (mit Transkripten):	259
Ereignisse mit Treffern:	253
Durchsuchte Sprecher (mit Transkripten):	689
Sprecher mit Treffern:	529

Types

Transkribierte Formen

gut (5899) ; besser (748) ; gute (333) ; guten (220) ; besten (189) ; guter (75) ; gutes (72) ; beste (62) ; bessere (37) ; jut (36) ; guad (21) ; besseren (20) ; beschte (16) ; gude (15) ; gud (13) ; gudd (12) ; besseres (11) ; guhut (11) ; gu (10) ; bester (10) ; gutt (8) ; guat (7) ; guder (7) ; gudi (6) ; beschten (6) ; bestes (5) ; best (5) ; gutem (4) ; besserer (3) ; beschde (3) ; bessre (3) ; guti (2) ; gutn (2) ; bessren (2) ; gudde (2) ; guddi (1) ; guth (1) ; guati (1) ; güd (1) ; gudes (1) ; khut (1) ; jute (1) ; jutet (1) ; besserem (1) ; guk (1) ; sputen (1) ; und (1) ; g (1) ; got (1) ; beter (1) ; jout (1) ; bestn (1) ; bezwei (1) ; beschden (1) ; gun (1) ; ju (1) ; guden (1) ; guts (1) ;

Normalisierte Formen

gut (6022) ; besser (750) ; gute (345) ; guten (239) ; besten (212) ; guter (82) ; gutes (75) ; beste (66) ; bessere (39) ; besseren (23) ; besseres (11) ; bester (10) ; best (5) ; gutem (4) ; besserer (3) ; Guten (3) ; Bestes (3) ; bestes (2) ; Gut (2) ; Guter (1) ; besserem (1) ;

Lemmatisierte Formen

gut (7898) ;

POS-Tags

ADJD (4001) ; NGIRR (2931) ; ADJA (966) ;

Figure 3: An excerpt of the lemma search of *gut* (good) showing the token frequency at each annotation level

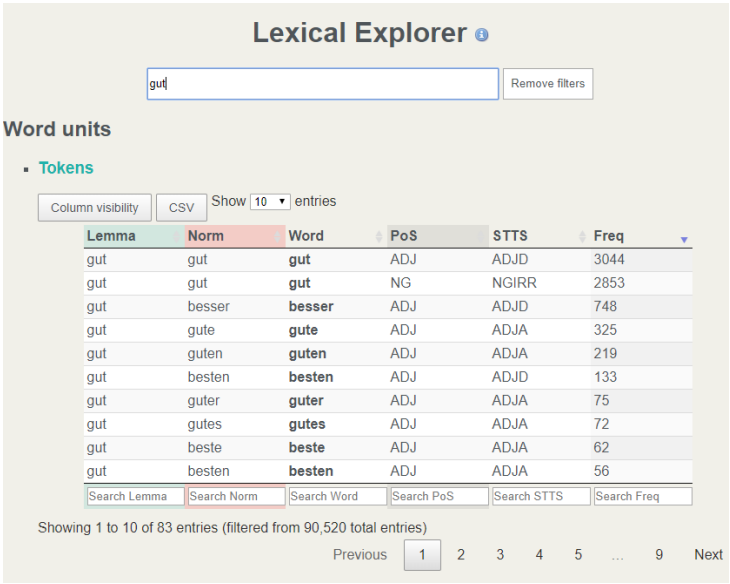


Figure 4: Default setting of columns for the bigrams table: the orthographic normalisations of the first (*Norm 1*) and second (*Norm 2*) bigram member and their absolute frequency (*Freq*)

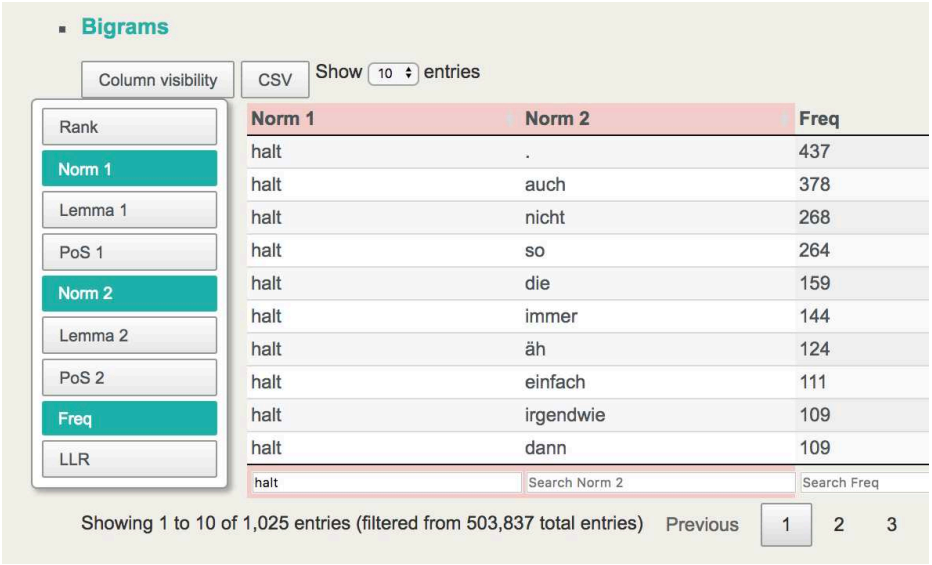
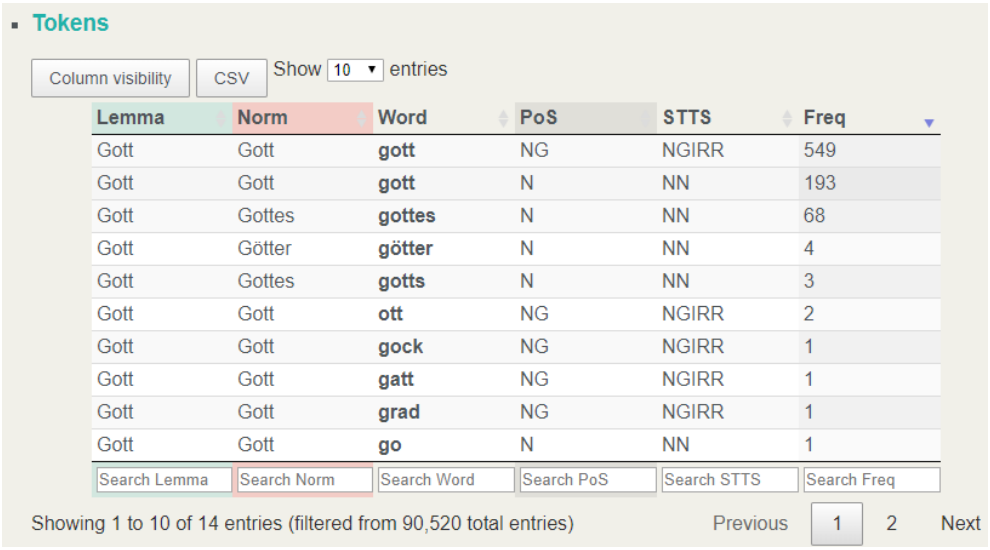


Figure 5: Occurrences of the word forms having the lemma *Gott* at each annotation level (*Freq*: frequency in FOLK)



▪ Tokens

Column visibility | CSV | Show 10 entries

Lemma	Norm	Word	PoS	STTS	Freq
Gott	Gott	**gott**	NG	NGIRR	549
Gott	Gott	**gott**	N	NN	193
Gott	Gottes	**gottes**	N	NN	68
Gott	Götter	**götter**	N	NN	4
Gott	Gottes	**gotts**	N	NN	3
Gott	Gott	**ott**	NG	NGIRR	2
Gott	Gott	**gock**	NG	NGIRR	1
Gott	Gott	**gatt**	NG	NGIRR	1
Gott	Gott	**grad**	NG	NGIRR	1
Gott	Gott	**go**	N	NN	1
Search Lemma	Search Norm	Search Word	Search PoS	Search STTS	Search Freq

Showing 1 to 10 of 14 entries (filtered from 90,520 total entries) Previous 1 2 Next

Figure 6: Excerpt from the table *Study corpus vs. DEREKO* (*HK Diff* stands for the difference of frequency classes

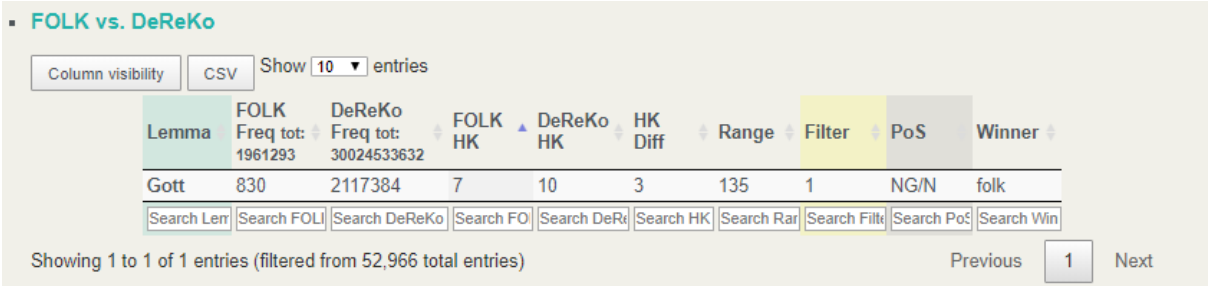
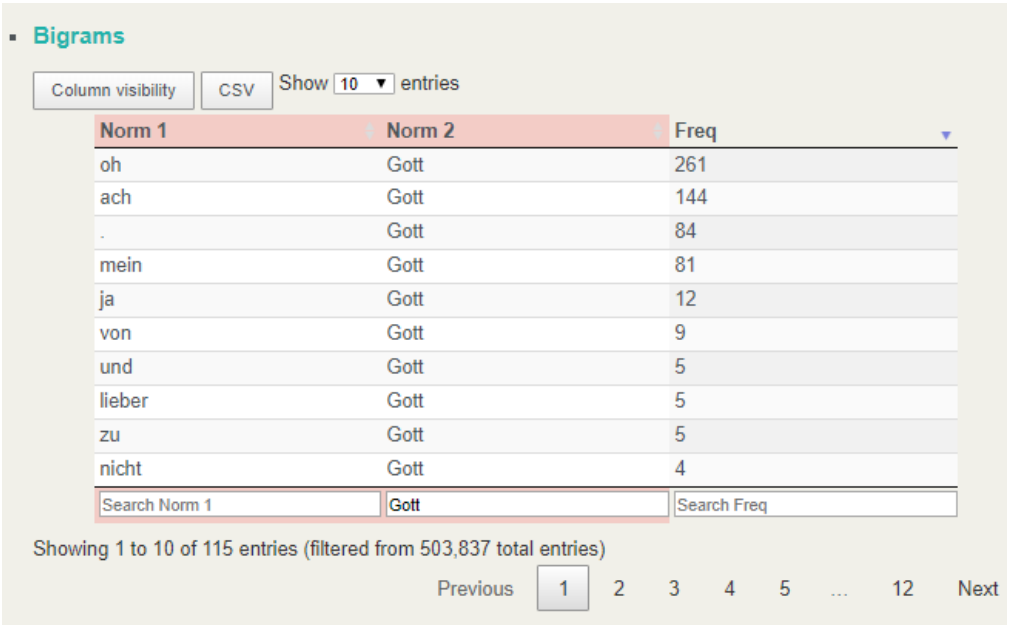


Figure 7: Excerpt from the table *Bigrams* showing the most common bigrams of *Gott* in FOLK in left context



Bigrams

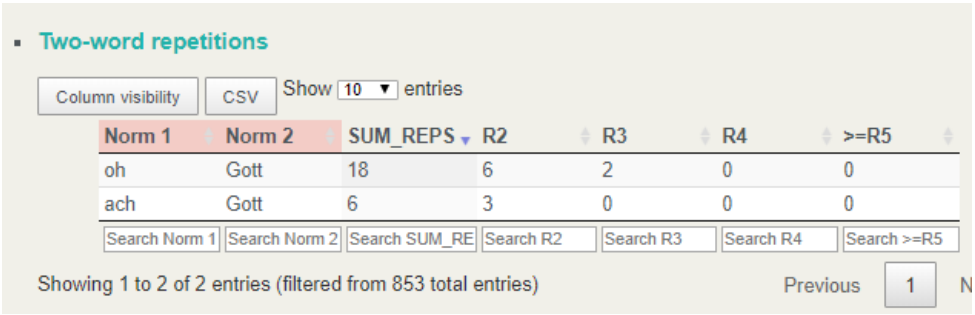
Column visibility CSV Show 10 entries

Norm 1	Norm 2	Freq
oh	Gott	261
ach	Gott	144
.	Gott	84
mein	Gott	81
ja	Gott	12
von	Gott	9
und	Gott	5
lieber	Gott	5
zu	Gott	5
nicht	Gott	4
Search Norm 1	Gott	Search Freq

Showing 1 to 10 of 115 entries (filtered from 503,837 total entries)

Previous 1 2 3 4 5 … 12 Next

Figure 8: Distribution of repetitions of *oh Gott* and *ach Gott* in FOLK: *SUM_REPS* refers to the sum of the repeated tokens; *R{x}* refers to the number of times the expression was repeated



- **Two-word repetitions**

Column visibility | CSV | Show 10 entries

Norm 1	Norm 2	SUM_REPS	R2	R3	R4	>=R5
oh	Gott	18	6	2	0	0
ach	Gott	6	3	0	0	0
Search Norm 1	Search Norm 2	Search SUM_RE	Search R2	Search R3	Search R4	Search >=R5

Showing 1 to 2 of 2 entries (filtered from 853 total entries) Previous 1 N

Figure 9: Excerpt from the table *Trigrams* showing *Gott* preceded and succeeded by a long stop <.>, standing for the contribution boundary

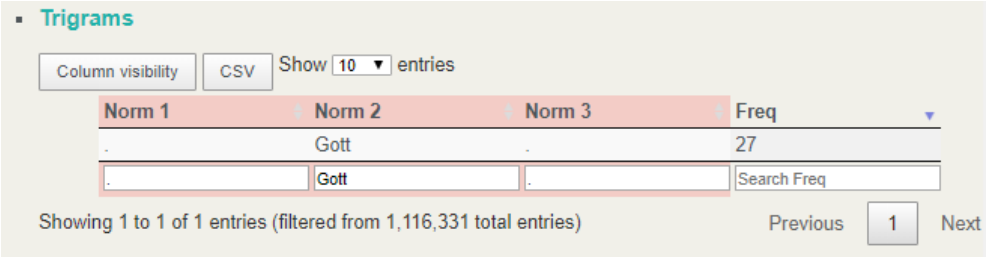
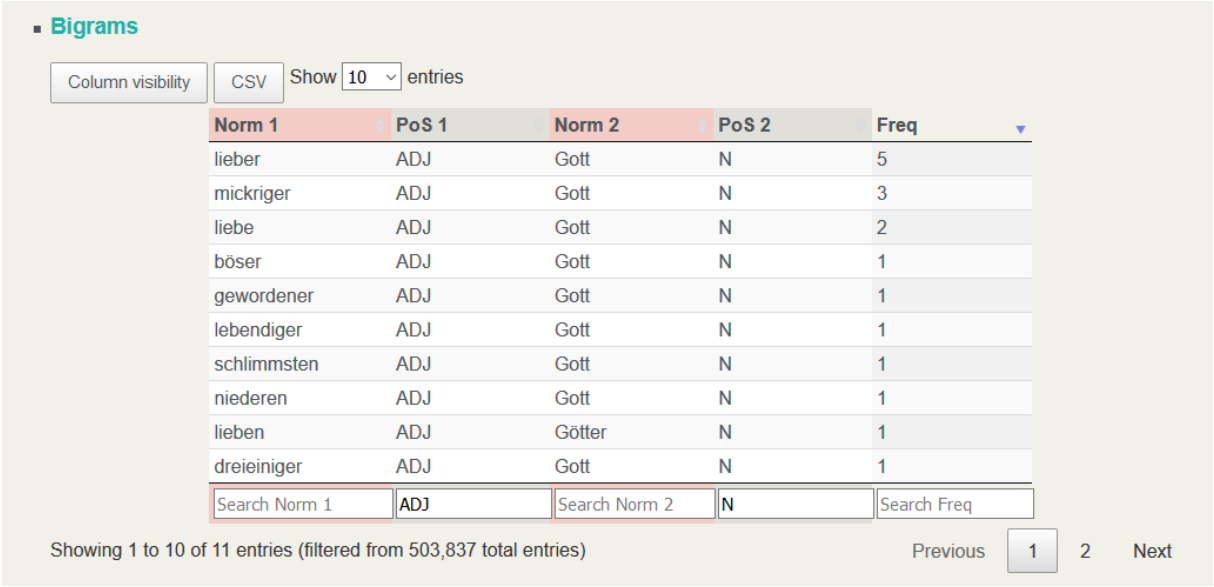


Figure 10: The most common bigrams with the word form *Gott* as second element if it is tagged as a noun (N) and if it is preceded by an adjective (ADJ)



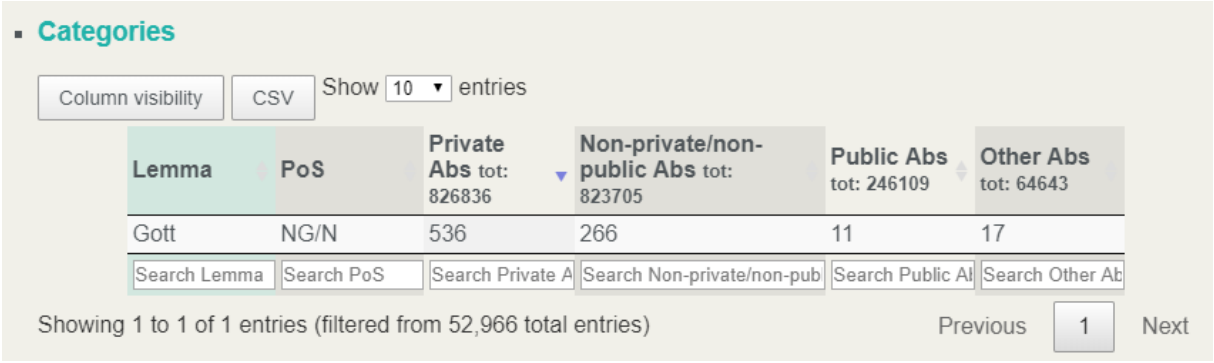
Bigrams

Column visibility CSV Show 10 entries

Norm 1	PoS 1	Norm 2	PoS 2	Freq
lieber	ADJ	Gott	N	5
mickriger	ADJ	Gott	N	3
liebe	ADJ	Gott	N	2
böser	ADJ	Gott	N	1
gewordener	ADJ	Gott	N	1
lebendiger	ADJ	Gott	N	1
schlimmsten	ADJ	Gott	N	1
niederen	ADJ	Gott	N	1
lieben	ADJ	Götter	N	1
dreieiniger	ADJ	Gott	N	1
Search Norm 1	ADJ	Search Norm 2	N	Search Freq

Showing 1 to 10 of 11 entries (filtered from 503,837 total entries) Previous 1 2 Next

Figure 11: Distribution of the lemma *Gott* in the categories private, non-private/non-public and public.



Tables

Table 1: Symbols for disfluencies in FOLK

Symbol	Meaning	Example
&	Abbreviations or word fragments caused by internal pauses	*Transcript:* em pe drei player *Normalisation:* MP3-Player & & &
%	Abruptions	*Transcript:* phonolo <pause> phonologischen *Normalisation* phonologischen phonologischen %
§	Idiolects, invented words	*Transcript:* in anderns portemonnaie *Normalisation* in § portemonnaie
#	Stutter	*Transcript:* lelele *Normalisation:* #
+	Unintelligeble	+++

Table 2: Self correction/abruption of *könnte* (could), FOLK_E_00069

Transcription	**könnt**	**könnte**	ich	mir	vorstellen
Normalisation	**könnte**	**könnte**	ich	mir	vorstellen

Table 3: Reduction in pronunciation of *mach mal* to *ma ma* in the contribution *ma ma scheiß ding au,* FOLK_E_00030

Transcription	**ma**	**ma**	scheiß	ding	au
Normalisation	**mach**	**mal**	scheiß	Ding	aus

Table 4: Excerpt of the table containing counts of exact one-word repetitions in FOLK

Lemma	Orth. normalisation	Simplified part-of-speech	Repeated (total tokens)	Repeated 2 times	Repeated 3 times	Repeated 4 times	Repeated >=5 times
äh	äh	NG	3010	1199	168	20	5
ja	ja	NG	2757	1028	141	40	21
ha	ha	NG	1742	266	147	80	74
nein	nein	NG	1540	438	102	48	28
he	he	NG	1241	301	117	39	25
ich	ich	P	962	425	34	1	1
und	und	KO	825	314	56	6	1
das	das	P	811	377	19	0	0
d	die	ART	503	250	1	0	0
da	da	ADV	494	193	22	4	5

Table 5: Tf-idf keywords of the event bible study group (FOLK_E_00193)

Lemma	Tf-idf
Sünde (sin)	0,00656
Jesus (Jesus)	0,00644
Geist (spirit)	0,00446
Liebe (love)	0,00425
Heilige (holy)	0,00422
hmhm (hmhm)	0,00411
lieben (to love)	0,00401
Salomo (Solomon)	0,00369
Gott (God)	0,00362
Gebot (commandment)	0,0034

Table 6: Wildcards for querying the Lexical Explorer

Wildcard	Function	Example
%	Matches zero or more characters	*ver%en* matches *versuchen, verwenden, verschwinden*, etc.
_	Matches one character	*_ehen* matches 1 character followed by *ehen* (e.g. *sehen*)
\| | Alternative search | *gehen\|laufen* matches *gehen* and *laufen* |
^ | Negative search (everything but x) | *^gehen* finds everything but *gehen* (it can be combined with alternative search) |

<
>
= | Numeric operators:
smaller then x,
bigger then x
same as x | >3 matches numbers higher than 3 |